Minihomework 1a

Roberto Navigli and Alessandro Scirè

SAPIENZA NLP

Entity Extraction

Goal: extract characters, places and organizations from the FairySum corpus; relate them within text

- **How:**
    - Apply Named Entity Recognition
    - **Extract:** people, locations, organizations
    - **Novelty:** work on connecting mentions:
        - We will request you to submit the **position** of each entity mention in the text and the **NER class**
        - Optionally provide a **unique string** identifying all the occurrences of an entity in the text
            - E.g. use “White Rabbit” (the longest? the first? your choice) to identify different occurrences, e.g., Rabbit, “White Rabbit”, etc.
        - Optionally, you will provide a link to an external knowledge resource (BabelNet or Wikidata, or any other resource)

What you will receive

- We provide you with a folder containing .txt files separated in two subdirectories, FAIRY_TALES and SHORT_STORIES
- Apply the NER tagging system of your choice to all the texts contained in the FAIRY_TALES/texts and SHORT_STORIES/texts directories
    - You can find the drive folder here!

What we expect from you (/1)

Generate a **.tsv** file with the following columns:

TEXT_ID: (string) corresponds to the filename without extension (e.g., bn:01899260n_The Raven)

START_OFFSET: (int) starting position of the span in the document

END_OFFSET: (int) ending position of the span in the document (inclusive, last character of the span)

LABEL: (string) NER category of the predicted span, from { PER, LOC, ORG }

NAME: (string, optional) unique name identifying all the occurrences of a given Named Entity in the text, e.g., “White Rabbit”

RESOURCE_ID: (string, optional) any identifier to look up the predicted Named Entity in an external resource, e.g., BabelNet or Wikidata ID.

- E.g. bn:03415301n, where bn:03415301n is the BabelNet ID for the “White Rabbit”

What we expect from you (/2)

- Submit a **Colab notebook** containing the code to execute the task
    - Make sure that the code can be executed without any error by running the command ‘Run all’ (Ctrl + F9)
- Write a **short report** (max 1 page) with the description of your work, highlighting the criticalities and the key insights
    - **Max 1 page**
    - **ACL 2021 paper template**
        - Available here (Overleaf LaTeX template) or here (Word and LaTeX direct download)
        - **DO NOT MODIFY** the template (margins, spacing, font size)
        - Use the non-anonymous flag, so you can enter your name

What we expect from you

- Submit a **.zip** file containing:
  - **results.tsv**, containing the predictions of the NER system
  - **ner.ipynb**, the **Colab notebook** containing the code to execute the NER tagging.
  - **report.pdf**, a short 1-page report describing what you did, issues, etc.
- Name the .zip file **`lastname_studentid_minihw1a.zip`**:
  - Ex: Luigi D'Andrea will submit a file named **`dandrea_1234567_minihw1a.zip`**
  - If you are unsure which name to put, use the one in your institutional email account
- For your submission, you must strictly use the following naming convention:

```
lastname_studentid_minihw1a.zip/
  -  results.tsv
  -  ner.ipynb
  -  report.pdf
```

⚠ Any violation to the above convention will invalidate your submission!

What you can and cannot do

- You can use whatever library (imported from your notebook) but a library that performs the expected task (i.e. NER ok, but not a library specific for character/location/organization extraction and aggregation)
- You **cannot** plagiarize!!!
  - If you do it, all students involved will be excluded

Submission

- You have to submit the .zip file through the submission form on Google Classroom

- You will be asked to fill a form with the requested information and attach the .zip file containing the results file (.tsv), the Colab Notebook and the report (.pdf).

- **Deadline: November 20th (AoE)**

Evaluation

Three elements of evaluation:

- We will evaluate the **predictions** provided in the **results.tsv** using a private evaluation script
- The **quality and novelty** of the code
- The **clarity** and **exhaustiveness** of the report you submit will be taken into account in your final score
- 33 is the maximum score (30 without any novelty / optional but being able to recognize entities "properly", e.g. including humanized animals)

Questions?

If you have a question that may interest your colleagues, **please ask it on Google Classroom**

Otherwise, for personal or other questions, email **ALL** of us (but please, only reach for things that can't be asked on the Google Classroom).

Our emails are:

{scire, orlando, bonomo}@diag.uniroma1.it

Good Luck!!!

Sapienza NLP